Exercises Analytics and Knowledge Discovery WS 2021

December 3, 2024

FH Salzburg
Informationstechnik &
System-Management

Contents

General rules

1. Recall from the syllabus that the exercise part must be positive ($> 50\%$) and can not be compensated by work afterwards! Tracking this is your responsibility as the teacher in the exercise only evaluates this after the last exercise.
2. Similarly, you need to be present in at least 75% of the exercise dates.
3. Conceptual exercises must be done without the computer and marked by the term "manual"
4. The applied exercises are formulated as questions by purpose because this mostly mimics the situation in real life. In practice nobody tells you the needed, singles steps as you are expected to be the expert. Therefore we also refrain from giving you more exact tasks.
5. Expect that we give you some other dataset/problem.
6. The goal is to discuss the evaluation and not the code. Why was the graphic/method chosen, what would be alternatives etc.. However, be prepared to answer questions about the code (in case something is not clear or does not work) or change the code.
7. Go through the lecture before preparing the exercise. You should get some hints WHAT to do.
8. Google only for hints HOW to do something in Python. Of course you are welcome to Google more information than were presented in the lecture but do not loose too much time.
9. ChatGPT is a tool that can be used according to the official guidelines in `https://myfhs.fh-salzburg.ac.at/display/REK/Einsatz+von+KI+in+Lehre+und+Studium+|+AI+in+Teaching+and+Studies`. But be aware: you need to show what was done and explain what was done and why it was done that way (and not possibly with another alternative). And be aware that ChatGPT can hallucinate and create nonsense output.

1 workflow and EDA

WF1 **Conceptual: data coding**. Consider the following raw data with customer reviews about a product. Code these data as a data matrix and describe how you coded the variables in a separate document.
Person 1: Larry Hagman from Pennsylvania, USA, red hair, age 76: "This software is amazing, does exactly what I need it to [Product Quality]. However, I do wish they'd stop raising prices every year as it's starting to get a little out of my budget."
Person 2: Melanie Firlinger, Kuchl, Austria, brown hair, age 43: "Love the product, but honestly I can't deal with the terrible customer service anymore. I'll be shopping around for a new solution."
Person 3: Franz Pirlinger, Hildesheim, Germany, brown hair, age 22: "Meh, this software is okay but cheaper competitors are just as good with much better customer service."

WF2 **Gathering data**: recommendation system
Consider the analysis done in `https://www.codespeedy.com/build-recommender-systems-with-movieler`

(a) Redo the analysis in your own notebook and understand what is done there. The analysis is simple without a complicated modeling approach. It is based on statistics, you should be familiar with correlation and confidence intervals. Think about the effects of some choices. Does the result make sense? Be prepared to do the analysis for a different movie.

WF3 **EDA** Pollution dataset: PM10 values in Lehen (fine dust, Feinstaub). Answer the following questions using proper graphics

(a) How many missing values do we have? If yes, can we just skip them?

(b) Did you see any unnormal PM10 values? How can you graphically assess this best?

(c) Show the distribution of PM10.

(d) How many times did PM10 exceed the limit in the whole year? With which standard plot can you assess this information graphically best?

(e) Do PM10 values have a relationship with the following variables: O3, PM10 in Rudolfsplatz, air temperature in Lehen and wind speed? Is the relationship linear or nonlinear?

(f) (*) Does PM10 vary throughout the day?

WF4 **EDA** Pollution dataset: ozone O_3. Answer the following questions using proper graphics

(a) How many missing values do we have? If yes, can we just skip them?

(b) Did you see any unnormal O3 values? How can you graphically assess this best?

(c) Show the distribution of O3.

(d) How many times did O3 exceed the limit in the whole year? With which standard plot can you assess this information graphically best?

(e) Do O3 values have a relationship with the following variables: NOX, air temperature in Lehen and wind speed? Is the relationship linear or nonlinear?

(f) (*) Does O3 vary throughout the day?

2 Data preparation

DP1 **Removal of missing data**: Consider the dataset dataIsRemovalAllowed which has the following three columns: gender (binary coded), age and preference score for a product. The goal is to assess the mean preference for a product (the preference is a score between 0 and 10).

(a) Can rows with missing data be removed without significantly influencing the result?
The answer is NO. Explain graphically, why nonrandom missingness of data here would lead to an underestimation of the preference score of the product.

DP2 **Data Imputation for timeseries data**
Load the dataset ex22PM10imputation containing timeseries of PM10 values of one year.
The first and the second column contain the same timeseries, but in the first column, 300 values are missing (entries Nan where Nan stands for Not a number). The goal is to impute 300 values using different methods and compare them to the real ones (in the second column).
Program and apply the following 4 methods for data imputation.
Hint1: the time interval is 30 minutes, so a day consists of 48 values.
Hint2: to make programming easier, values are only missing one at a time, i.e. there are no missing values following each other.

(a) Substitute the Nan values by taking the **last known value**.

(b) Substitute by **linear interpolation** using the last known and the next known value.

(c) Substitute with the **value of the day before** at the same time.

(d) Substitute with the **mean value of the whole time series** (generic method not only applicable for timeseries).

DP3 **Discussion of the data imputation result using EDA**
Describe the results of the data imputation example before using EDA. Answer the following questions using appropriate graphics.
An important, recurring topic in data analysis is evaluation, especially computation and description of errors: besides choosing suitable graphics (suitable for the question to be answered) one also needs to choose a suitable error measure and often also a reference for sake of comparison.

(a) Which of the 4 data imputation methods is best suited for this timeseries? For the following 3 questions choose this method.

(b) Is the imputation error high? Can it be expected that imputation does not harm the following analysis?

(c) What is the distribution of the imputation error? What is the probability, that the relative error is more than for example 5%? (Hint: you will probably need 2 graphics here)

(d) Plot the whole timeseries and add the imputed data points. We will use this plot to discuss together why this result is not unexpected for this special timeseries and if imputation is expected to harm following analyses or not.

DP4 **Data Imputation with regression** based on other variables

(a) Apply data imputation to the missing PM10 values using a properly done linear regression with (a subset of) other, known variables. Take care to properly choose the(se) variable(s) for example based on your EDA in exercise 1.4.

(b) As we do not know the real value, no imputation error can be computed. But we can compare methods. Therefore also do the imputation with the last value. Then compare the imputed values using the regression-method with the imputed values using the last-value-method by making a suitable graphic.

DP5 **Feature selection** in Python
Look at `https://machinelearningmastery.com/feature-selection-for-regression-data/`
Redo the first part of the analysis done there:

(a) Run the code using features selected based on correlation of the feature with the outcome and explain what is done there.

(b) Run the code using features selected based on the mutual information of the feature and the outcome and explain what is done there.

(c) The important decision here is to **select the number of features k**. Do this **manually** for both methods by running the code with k= 5, 10, 20, 30, 40, 50, 60, 70, 80, 90. Then draw the resulting errors depending on k.
Which of the two methods seems to be better? Which k should we select?

DP6 **Feature generation** for linear regression
Load the data dataFeatureGenerationForLinReg.csv. The first two columns are the input variables x_1 and x_2, the third is the outcome y which should be predicted from x_1 and x_2.

(a) Predict the outcome y from x_1 and x_2 using a purely linear model and assess the fit graphically.

(b) Predict the outcome y from x_1 and x_2 but now add polynomial terms to the linear model. Again assess the fit graphically.

3 Principal Components Analysis (PCA)

PC1 Apply the PCA to the superhero-data.
Hint: For a sample analysis in Python see `https://www.reneshbedre.com/blog/principal-component-ana` `html` Look at the iris example in the lower part.

(a) Calculate the PCA and give out the principal components. Should one normalize before? Show the coeffients of the first principal component.

(b) Assess the number of needed dimensions

(c) Interpretation of the transformation: show the biplot with labeled variable names and describe, what you see there: what do the first 2 PC's mean?

PC2 PCA for MNIST digit data.

(a) Calculate the PCA. Should one normalize before?

(b) Assess the number of needed dimensions

(c) Visualize the transformed data. Can the different digits be separated?

PC3 Movie data

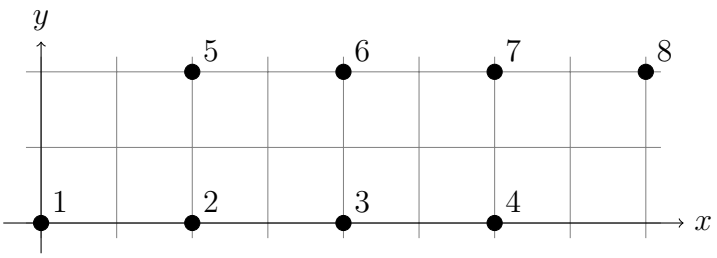
(a) Load and very briefly describe the movie data (data understanding), i.e. what rows and columns do we have, what entries are there?

(b) Apply a PCA to the movie data. How many dimensions should we keep in order to not loose too much information?

(c) Show the biplot with labeled variable names and describe, what you see there. How useful is this plot? Why or why not (depending on your answer before)?

4 k-Means clustering

KM1 Conceptual: Consider the following data points.



Consider the following 3 possible clustering results

- Clustering 1: $C_1 = \{1, 2, 3, 4\}$, $C_2 = \{5, 6, 7, 8\}$
- Clustering 2: $C_1 = \{1, 2, 5\}$, $C_2 = \{3, 4, 6, 7, 8\}$
- Clustering 3: $C_1 = \{1, 2, 5\}$, $C_2 = \{3, 6\}$, $C_3 = \{4, 7\}$, $C_4 = 8$

(a) Determine **manually**, which of the clustering results given above is favored by kmeans' loss function. The question posed here is not really correct. Why?

(b) Consider the following 3 cluster centers: $\vec{c}_1 = (0, 1), \vec{c}_2 = (5, 0), \vec{c}_3 = (6, 2)$. Sketch, how the points would be assigned to the 3 cluster given these cluster centers (i.e., plot the Voronoi-diagram).

KM2 Illustrative synthetic dataset, determination of k, description of clusters.

(a) Cluster the balls data using kmeans.
Try to choose k based on data visualization
Try to choose the number of clusters based on finding the elbow of the loss function dependent on k
Output the cluster centers
Assess how many samples are contained in each cluster.

KM3 Illustrative synthetic datasets, determination of k, description of clusters.

(a) Cluster the two Moons data using kmeans. Choose k both with the generic elbow method and by visualization. Output the cluster centers. Is the result good?

(b) Cluster the circles data using kmeans. Choose k both with the generic elbow method and by visualization. Output the cluster centers. Is the result good?

KM4 Movies dataset

(a) Cluster the movies data using kmeans. Try to find a good k

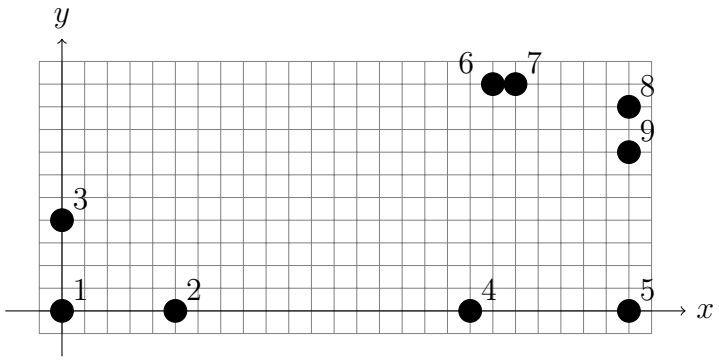
(b) For sake of simplicity here we set $k = 5$. Output the cluster centers and use it to interpret the meaning and sizes of the found clusters.

(c) (*) Analyze the meaning of the 5 clusters by visualizing each cluster using a WordCloud.

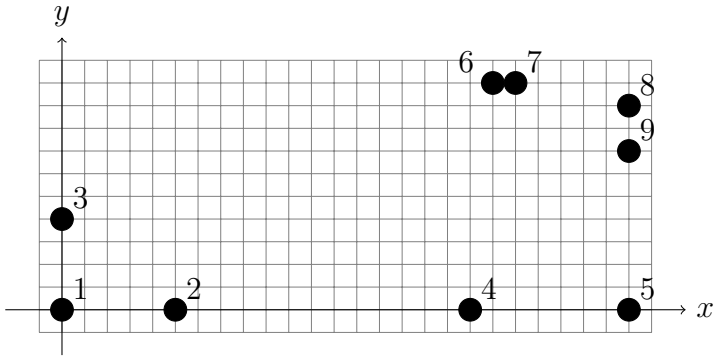
5 Hierarchical clustering

HC1 Conceptual: consider the following data points.

(a) Consider Manhattan distance (one grid denoting here 1 unit for simplicity) and single linkage. In which order are the points/ clusters clustered together? Show this by drawing circles around (as it was done on the slides).
Manually construct and plot the corresponding dendogram without the need to calculate the correct heights. In case of several next possibilities choose one.

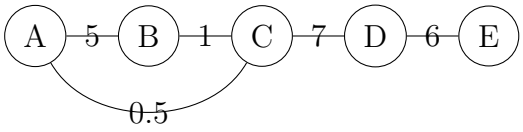


(b) Do the same for **complete** linkage. The image is repeated so that you can use it to illustrate the way the algorithm constructs the dendrogram.

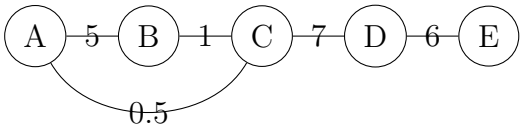


HC2 Conceptual: hierarchical clustering for graphs

(a) determine 2 clusters of nodes of the following graph using hierarchical clustering using single linkage. Could you cluster this graph using k-means clustering, too?



(b) Do the same using complete linkage. Hint: lines that are not drawn correspond to similarity zero. Do you find the same clusters? Again, for your convenience, the graph is repeated.



HC3 Illustrative, synthetic datasets

(a) Successfully cluster the balls data using hierarchical clustering.
Plot the dendrogram: how many clusters would be reasonable by its inspection?
Try out different linkage options: do all of them work well?

(b) Successfully cluster the twoMoons data using hierarchical clustering.
Plot the dendrogram: how many clusters would be reasonable by its inspection?
Try out different linkage options: do all of them work well?

HC4 Movies dataset

(a) Cluster the movies dataset with hierarchical clustering choosing a suitable linkage function
Plot the dendrogram: how many clusters would be reasonable by its inspection?

(b) Interpretation of the clustering result: describe the found clusters as for the corresponding kmeans example and try to find out the differences (use 5 clusters again to enable the comparison).

6 t-distributed Stochastic Neighborhood Embedding (tSNE)

tSNE1 Conceptual: Input similarities
Load the data dedicated for this and the next example from Moodle. Plot the data.

(a) Program the calculation of the similarity S in the input space using a constant σ_i for all i as described in the lecture.

(b) Change the value of σ_i: describe the effect on the similarities (of course graphically).

tSNE2 (*) Conceptual: Algorithmic determination of the local scale.
Consider the same data as before and the results from exercise 1.

(a) Write a program that determines the entropy of the computed similarity values.

(b) Write a binary search that determines the local scale σ_i for any specified single data point.

tSNE3 (*) Conceptual: Gradient descent in 1-D.
Consider the following loss function

$$\text{loss}(x) = x^5 - 5x^4 + 5x^3 + 5x^2 - 6x$$

(a) Program the one-dimensional gradient descent algorithm (using the exact derivative).

(b) The loss function has two local minima. Find them using your gradient descent function. Experimentally determine the set of starting points x_0 that lead to the first local minimum and the set of starting points x_0 that lead to the second local minimum.

tSNE4 Balls data

(a) "Reduce" the dimension of the balls data to 2 dimensions using tSNE. Plot the result.

(b) Reduce the dimension of the balls data to 1 dimension and plot the result.

(c) Vary the perplexity such that it is too low and too high, respectively. What happens when the perplexity is set too low? What happens when the perplexity is set too high?

tSNE5 MNIST data

(a) Study the convergence of the algorithm:
Set the perplexity to 25 and try the following values for the maximum number of iterations: 50, 100, 200, 400, 750, 1000. Visualize the result for each choice. Is the standard setting (which is 1000) enough? Can one assess convergence based on a single run?

(b) Perplexity determination: Try several values. Visualize the results and choose one.

(c) Assess stability: Perform 4 runs with the chosen perplexity.
How stable is the result? Why does one get different results for each run?

tSNE6 Movies data

(a) Perplexity determination: Try several values. Visualize the results and choose one.

(b) Cluster your reduced data using kmeans with a good selection of k.
Describe the resulting cluster centers. Can you transform them back to the original space?

7 Spectral clustering

SC1 Conceptual: similarity matrix, similarity graph and nCut
Consider the following data points, where 1 tick is 1 unit wide. **Do this example manually, i.e. with pen and paper**

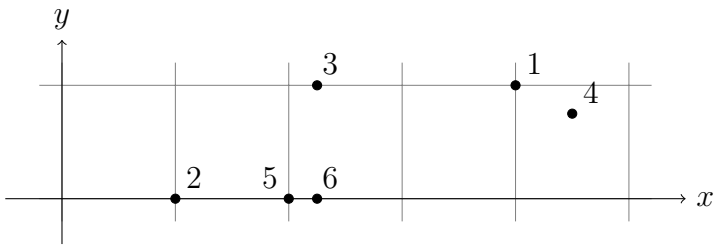


Figure 1: Data points for spectral clustering calculations

(a) Determine the similarity-matrices with the setting and nNeighbors=2. Draw the corresponding similarity graph (can also be drawn in the figure above).

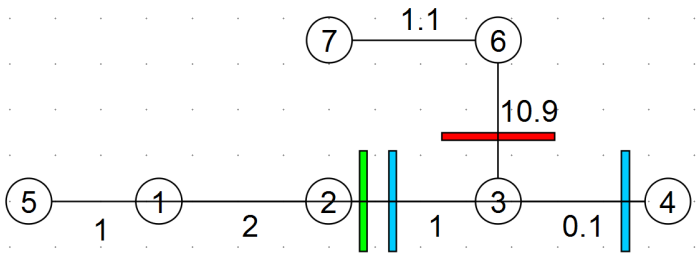
(b) Determine the similarity-matrices with the setting and nNeighbors=4. Is this a good setting?

(c) Determine the corresponding nCut-criterion resulting when cut the similarity graph such that you form the clusters {2,5,6,3} and {1,4}.
Can you find a better clustering that can be achieved according to the nCut-criterion?

SC2 Conceptual (Algorithm, part Graph Laplacian matrix)
For the points above and nNeighbors=2 determine the resulting normalized graph Laplacian L_{rw} **manually**.

SC3 Conceptual (Algorithm, last part)
Use the resulting normalized graph Laplacian L_{rw} from the last example. Program the remaining part of the algorithm and apply it to L_{rw} to get the 2 clusters.

SC4 Conceptual: nCut
Manually calculate the Ncut performance criterion for the 3 clusterings (defined by the differently coloured cuts) of the following undirected, weighted graph. Based on that determine which of these three clusterings is the best one.



SC5 Practical Clustering.

(a) Cluster the **circles** data successfully .

(b) Cluster the **face** data successfully.

(c) Cluster the **two moons** data successfully.